Adaptation Data Selection using Neural Language Models:
Experiments in Machine Translation

Kevin Duh
Graham Neubig

Katsuhito Sudoh
Hajime Tsukada

NAIST
NTT
The Big Picture

IN-DOMAIN BITEXT

Data Selection using Language Models (LMs)

GENERAL-DOMAIN BITEXT

SUBSET

Many unknown words here! Are Ngram LMs sufficient?

Training Pipeline

MT SYSTEM
Research Question:

– To investigate alternative LMs for data selection
– In particular: Neural LM
  • Their continuous word representations have been shown to be robust to unknown words

Result:

– Data selection by Neural LMs improve over Ngram LMs by \textbf{0.1-1.7 BLEU} (TED Talks tasks).
Data Selection Criteria

[A. Axelrod, X. He, J. Gao. Domain Adaptation via Pseudo In-Domain Data Selection (EMNLP11)]

1. Score each General-Domain sentence-pair (e,f) by 4 LMs:

\[
\text{[CrossEntropy}(\text{LM}_{\text{IN},e}^e) - \text{CrossEntropy}(\text{LM}_{\text{GENERAL},e}^e)]
\]
\[
+ \ [\text{CrossEntropy}(\text{LM}_{\text{IN},f}^f) - \text{CrossEntropy}(\text{LM}_{\text{GENERAL},f}^f)]
\]

*Preference:

- Prefer sentences similar to in-domain bitext
- Prefer sentences dissimilar to average general domain bitext

2. Rank sentence pairs by score; threshold by validation set
Ngram vs. Recurrent Neural LM

\[ P(w_t | w_{t-1}, w_{t-2}) \]

*Backoff is needed for rare or unknown contexts \((w_{t-1}, w_{t-2})\)*

e.g.

“recite Shakespeare’s poem”
\[ \rightarrow P(\text{poem} | \text{Shakespeare’s recite}) \]

“recite ScoobyDoo’s poem”
\[ \rightarrow P(\text{poem} | \text{ScoobyDoo’s recite}) \]
\[ \rightarrow P(\text{poem} | \text{ScoobyDoo’s}) \]
\[ \rightarrow P(\text{poem}) \]

[T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, S. Khudanpur. Extensions of Recurrent Neural Net Language Model (ICASSP11)]

*Continuous vector representations enable sharing of context!*

Previous Word \(w_{t-1}\)

Previous State \(s_{t-1}\)

\[ \text{sigmoid}(Uw_{t-1} + As_{t-1}) \]

Current State \(s_t\)

Current Word \(w_t\)

\[ \text{softmax}(Vs_t) \]
Experimental Setup: 4 language-pairs

IN-DOMAIN:
TED Talks from WIT³ (~130k sentences)

GENERAL-DOMAIN:
All WMT2013 data, e.g. Europarl, News-crawl (2M-40M sentences)

1. Data Selection using Ngram vs. Neural LMs

2. Standard Moses Pipeline

3. Determine best subset size {10k, 50k, 100k, 500k, 1M} sentences by validation set BLEU

4. Compare test set BLEU
Analysis

1. Are improvements due to lower OOV rate or better estimates of translation probability?
   – Force decoding gives similar BLEU → better estimates

2. How much overlap between sentences selected by Ngram vs. Neural LM?
   – 60-75% overlap, so this is incremental improvement

3. Computation time?
   - Fast to train Neural LMs for small in-domain set
Summary of paper in Haiku
(thanks to Chris Quirk for poetic inspiration)

These Neural LMs

Easy and good like N-grams

Why don’t you try them?

Code/Scripts available: http://cl.naist.jp/~kevinduh/a/acl2013/